

로지스틱 회귀 분석을 통한 보험 사기 예측

TEAM 박상희 이인풍 박서희 박소영

2018.12.10

- 001 분석 배경 및 자료 설명
- 002 EDA
- 003 로지스틱 회귀 분석
- 004 예측 및 결론

Part 1.

분석 배경 및 자료 설명



- **분석 배경**



- **데이터 분석 팀에 발령!!!**
- **첫 분석 업무가 주어짐!!**
- **최근 보험사기를 통해 보험금을 수령하는 경우가 늘어나고 있음.**
- **보험 사기인지 아닌지 예측하는 알고리즘 개발**

- **어느 보험회사에 취직한 신입사원 4명**

• 자료 출처



- 2016년 빅콘테스트에서 제공한 자료
- 최근 10여년 동안의 보험사기자와 일반 고객으로 구분하여 구성
- 고객정보, 계약정보, 지급정보, 설계사 정보 등이 있음
- 이 중 분석에 필요한 데이터만 따로 추출하여 분석 진행

• 데이터 샘플링

- 1단계 6,299 obs 와 16 Variables 로 데이터 생성
- 2단계 한 개의 변수라도 NA인 데이터 3,849개 제거 → 2,450 obs
- 3단계 일반고객(2,172 obs)과 보험사기(278 obs)의 데이터 중 8:2 의 비율로 샘플링
- 4단계 일반고객(800 obs)과 보험사기(200 obs)의 데이터 최종 구성

• 데이터 분리

- 1단계 1,000 obs 중 80% : 20% 의 비율로 Train data 와 Test data 분리
- 2단계 Train data = 일반고객(640 obs) + 보험사기(160 obs)
- 3단계 Test data = 일반고객(160 obs) + 보험사기(40 obs)
- 4단계 Train data로 모형을 만든 후 Test data로 예측 진행

• 변수 설명

• 반응변수



YES : 보험 사기

NO : 일반 고객

• 범주형 설명 변수 (7)

- 진단서 상의 병원의 종류
- 직업
- 사고 원인
- 보험 청구 횟수
- 결혼 유무
- 보험설계 경력
- 입원으로 보험 청구한 수

• 연속형 설명 변수 (8)

- 나이
- 보험 담당자 변경 횟수
- 보험 가입 수
- 보험료 총 납입액
- 고객 추정 소득
- 계약 후 보험 청구까지의 최소 기간
- 계약 후 보험 청구까지의 최대 기간
- 보험 해약 비율

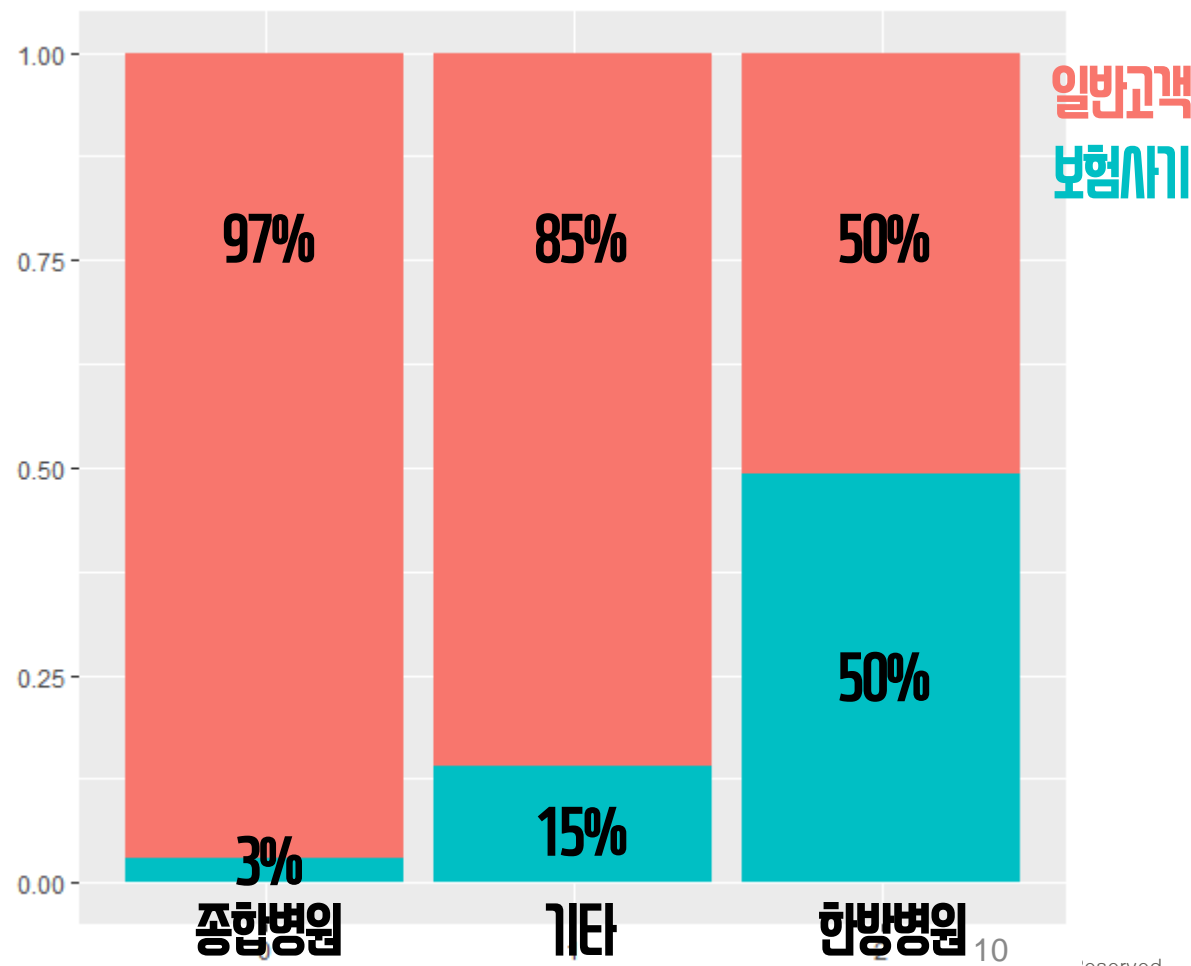
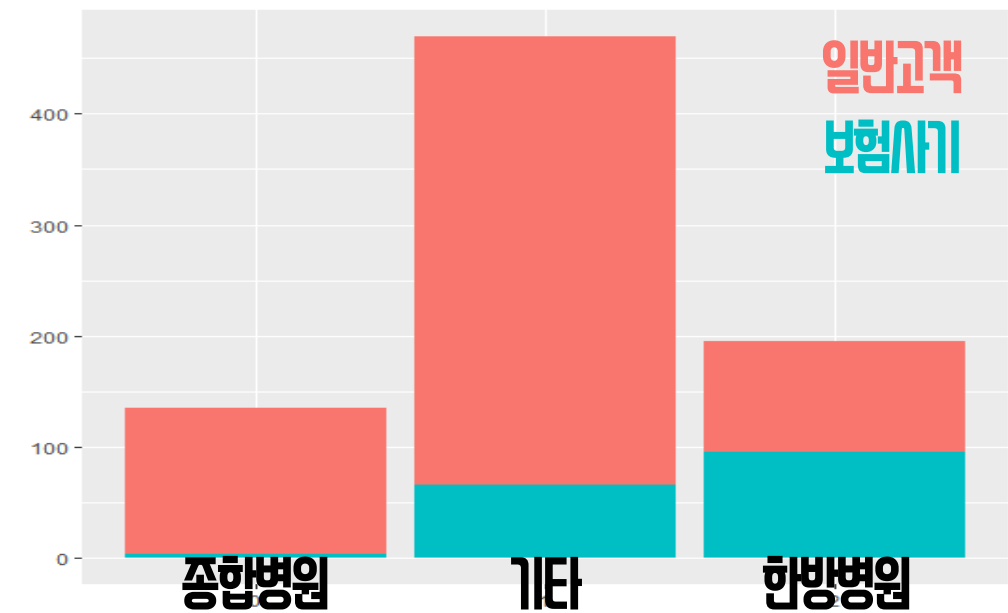
Part 2.

EDA



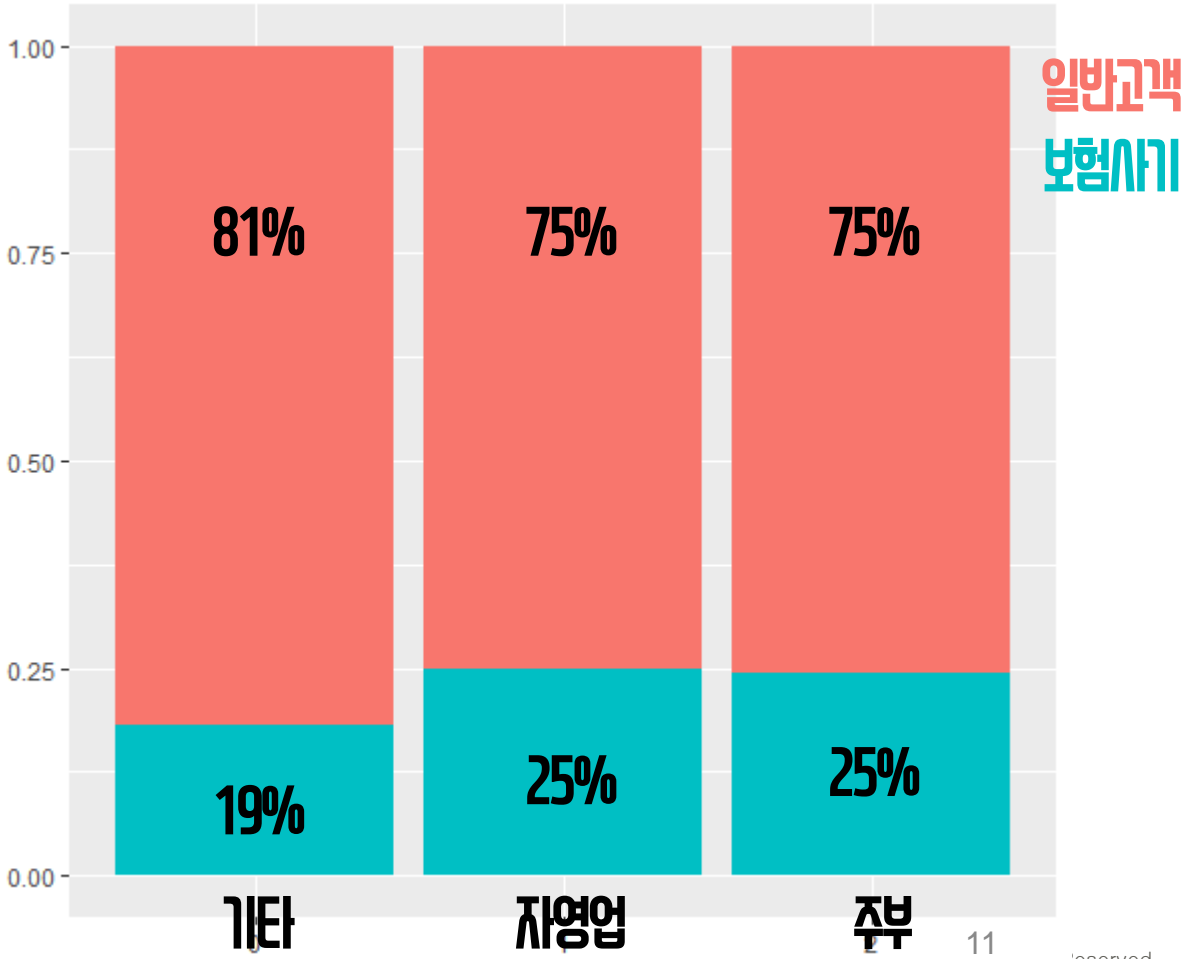
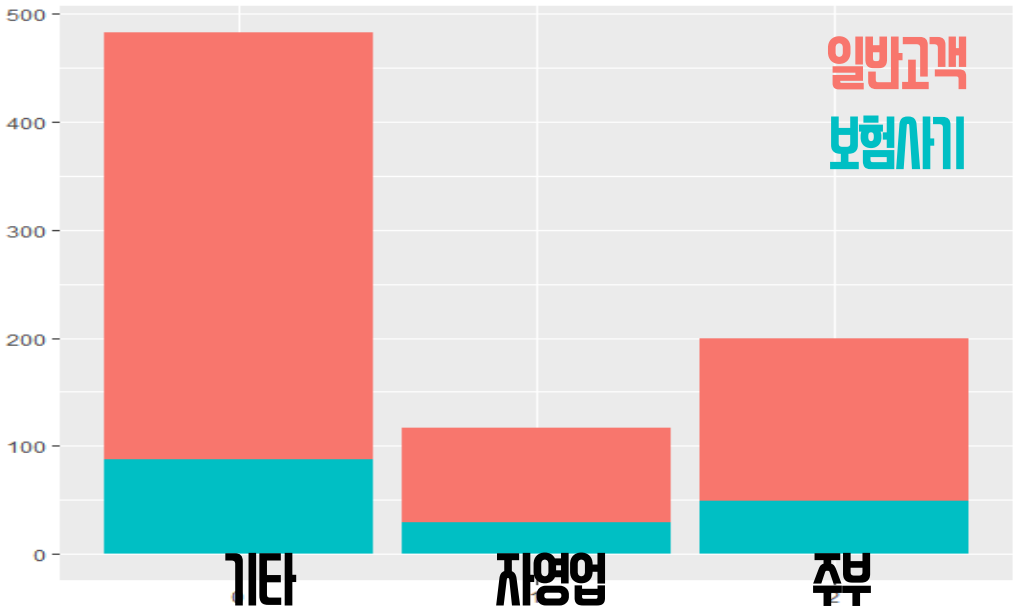
진단서 상의 병원의 종류

- 보험을 청구할 때 진단서를 같이 제출
- 범주 구성 : 종합병원, 한방병원, 기타



• **고객의 직업**

- **범주 구성 : 기타, 주부, 자영업**
- **주부와 자영업자의 보험사기자 비율이 높았다.**

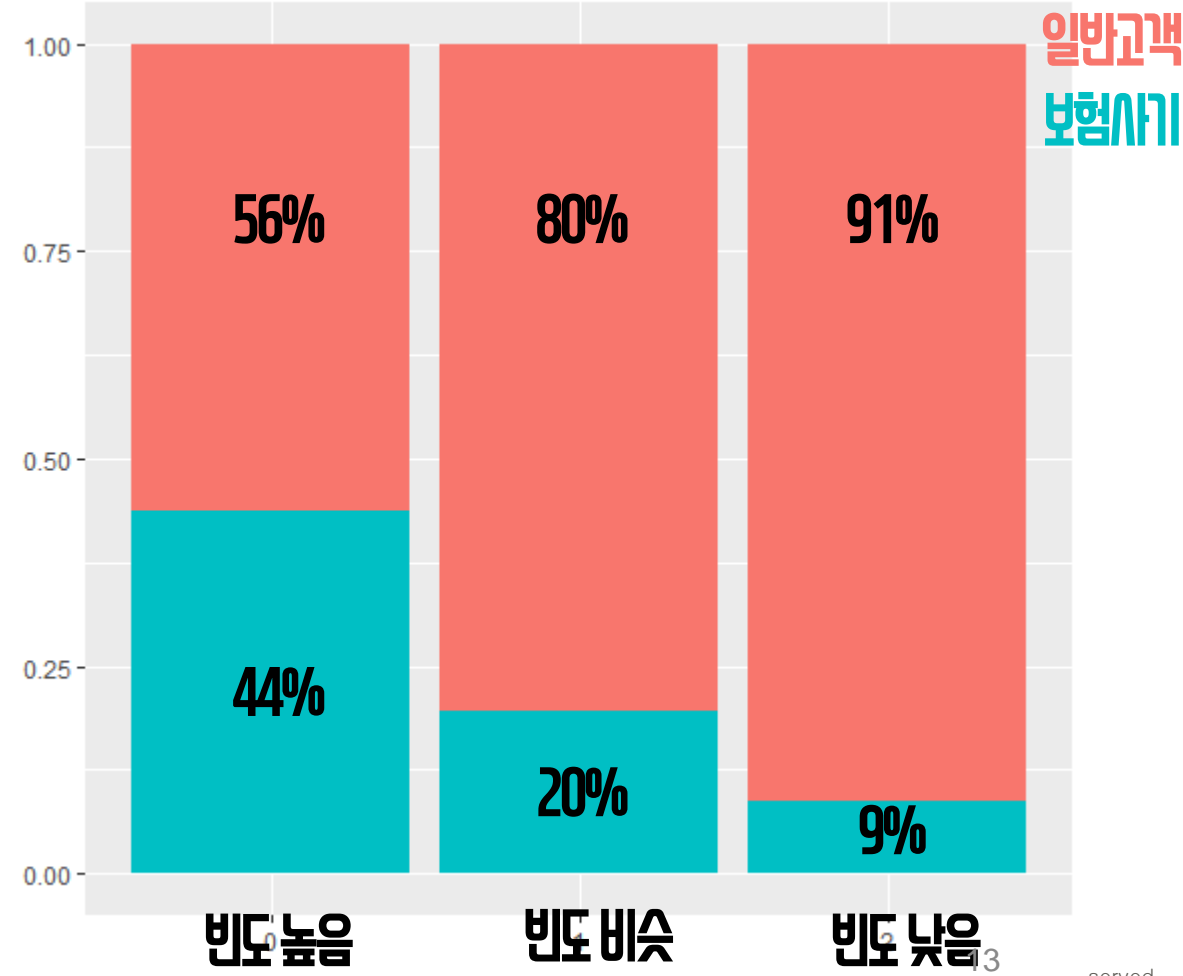
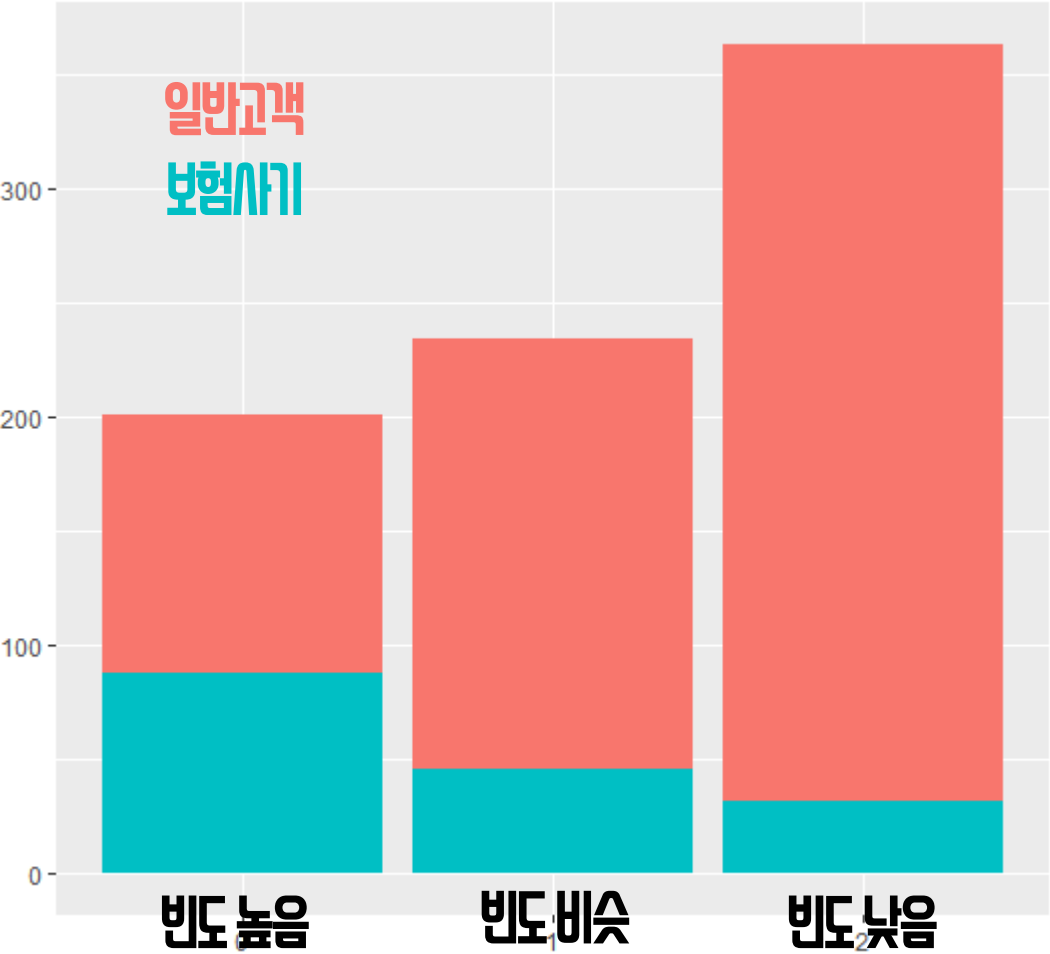


• 사고 원인

- 보험 업계에서는 **질병 분류 코드**를 317개의 코드로 구분
- 각 317개의 코드에 대하여 보험사기자와 일반 고객의 비율을 구분
- 범주 구분 : 보험사기빈도가 높은 질병, 보험사기빈도가 비슷한 질병, 보험사기빈도가 낮은 질병
- ex. 보험사기빈도가 높은 질병 : **골절, 골반 통증**
- ex. 보험사기빈도가 비슷한 질병 : **알레르기, 복부 통증**
- ex. 보험사기빈도가 낮은 질병 : **유방암, 당뇨병, 고혈압**

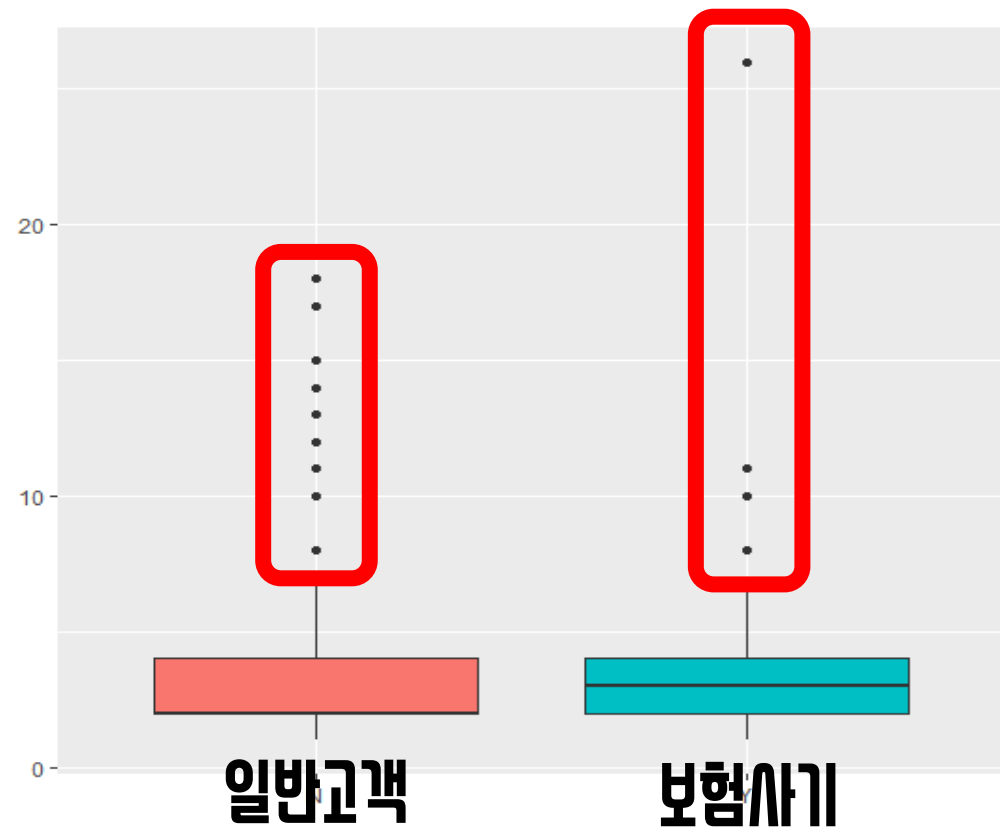
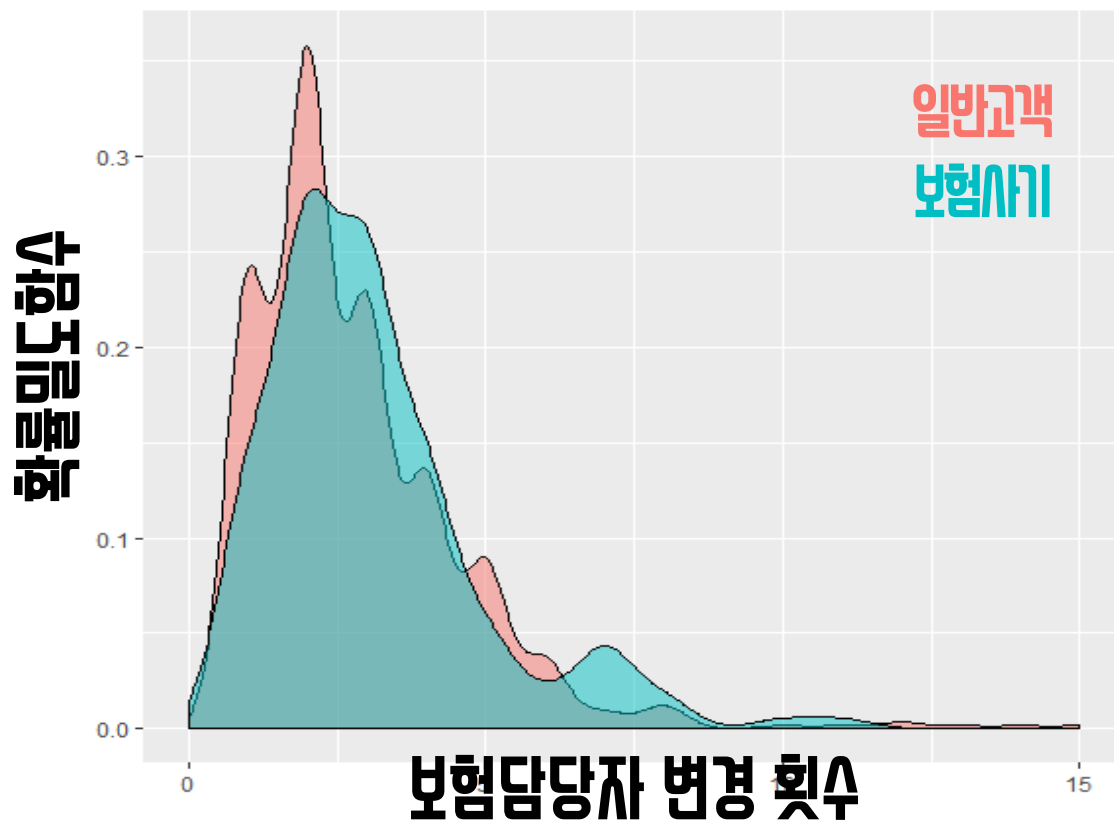


• 사고 원인



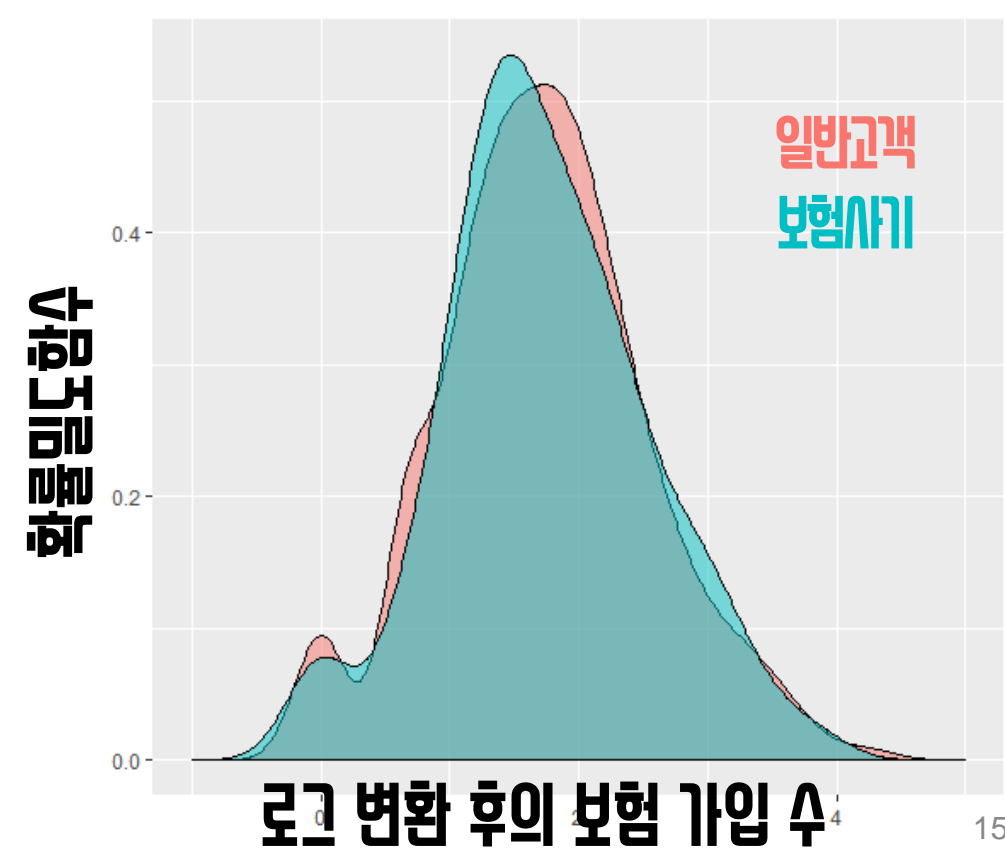
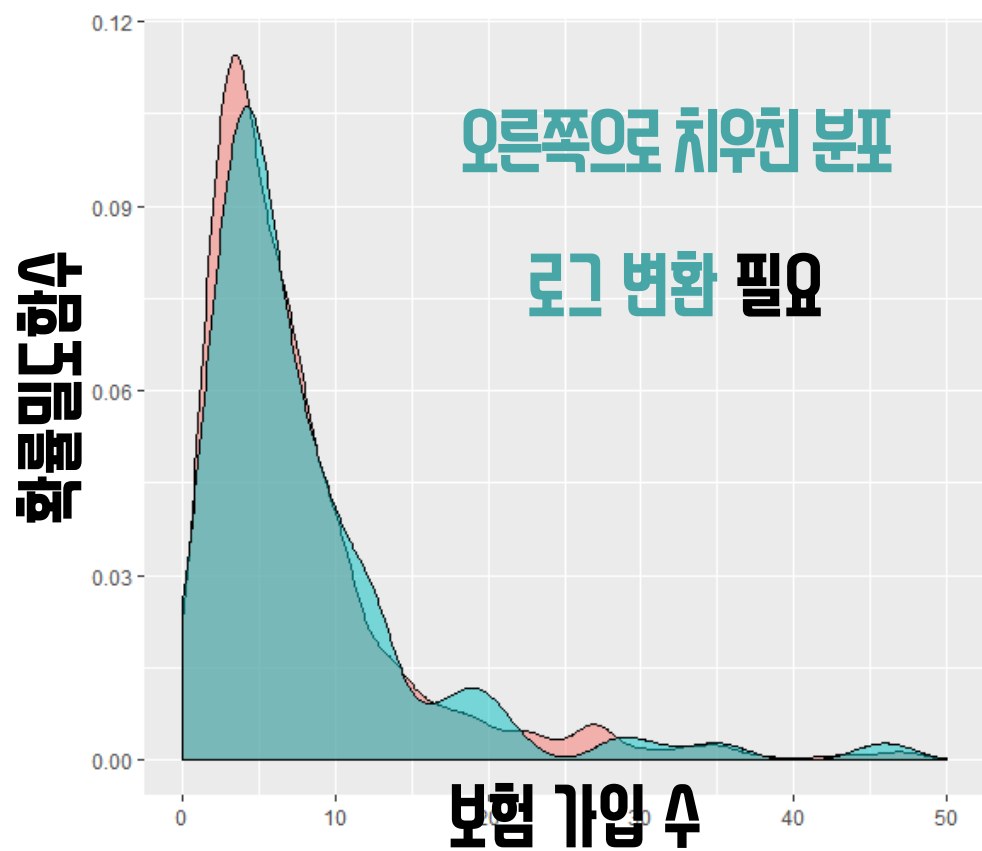
- 보험담당자의 잦은 변경

- 보험담당자를 자주 변경할 경우 보험사기일 가능성이 높다고 생각.



• 보험 가입 수

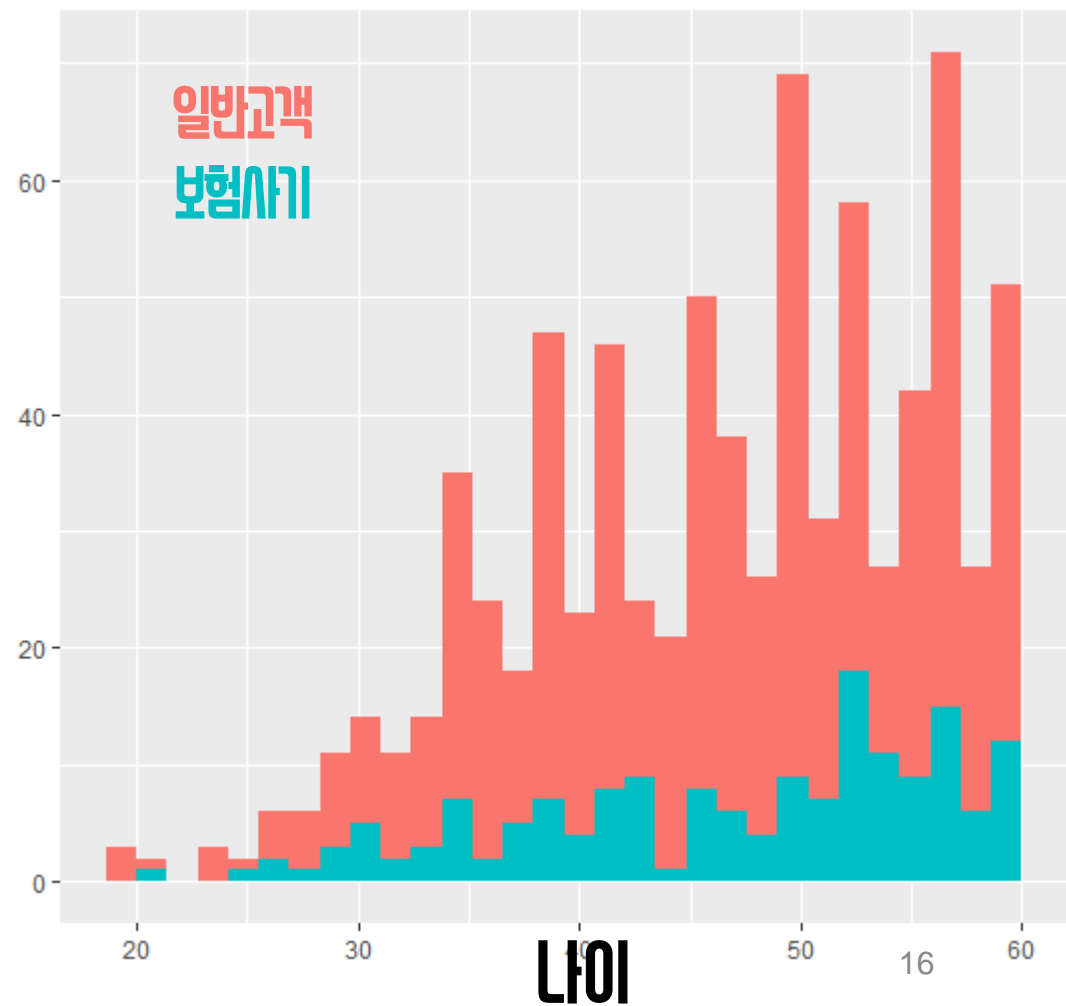
- 일반적으로는 보험을 10개 미만을 가입하지만, 10개 이상의 많은 보험을 가입하는 경우도 적지 않았다.



- 가입 고객의 나이

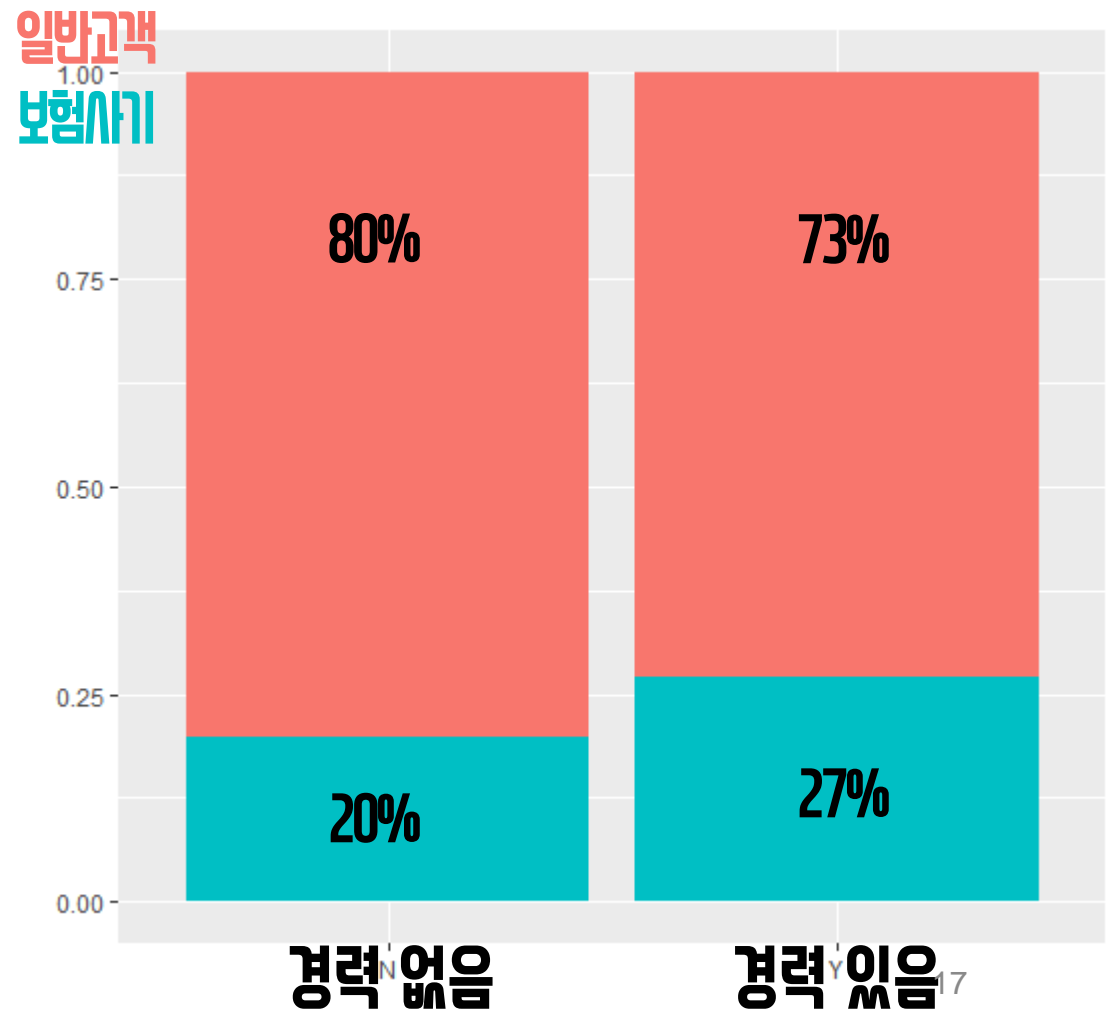
- 나이가 많을수록 가입 고객의 수 ↑

- 일반고객과 보험사기자 나이의 분포 비슷
(일반고객의 평균 나이 : 46세 / 보험사기의 평균 나이 : 47세)



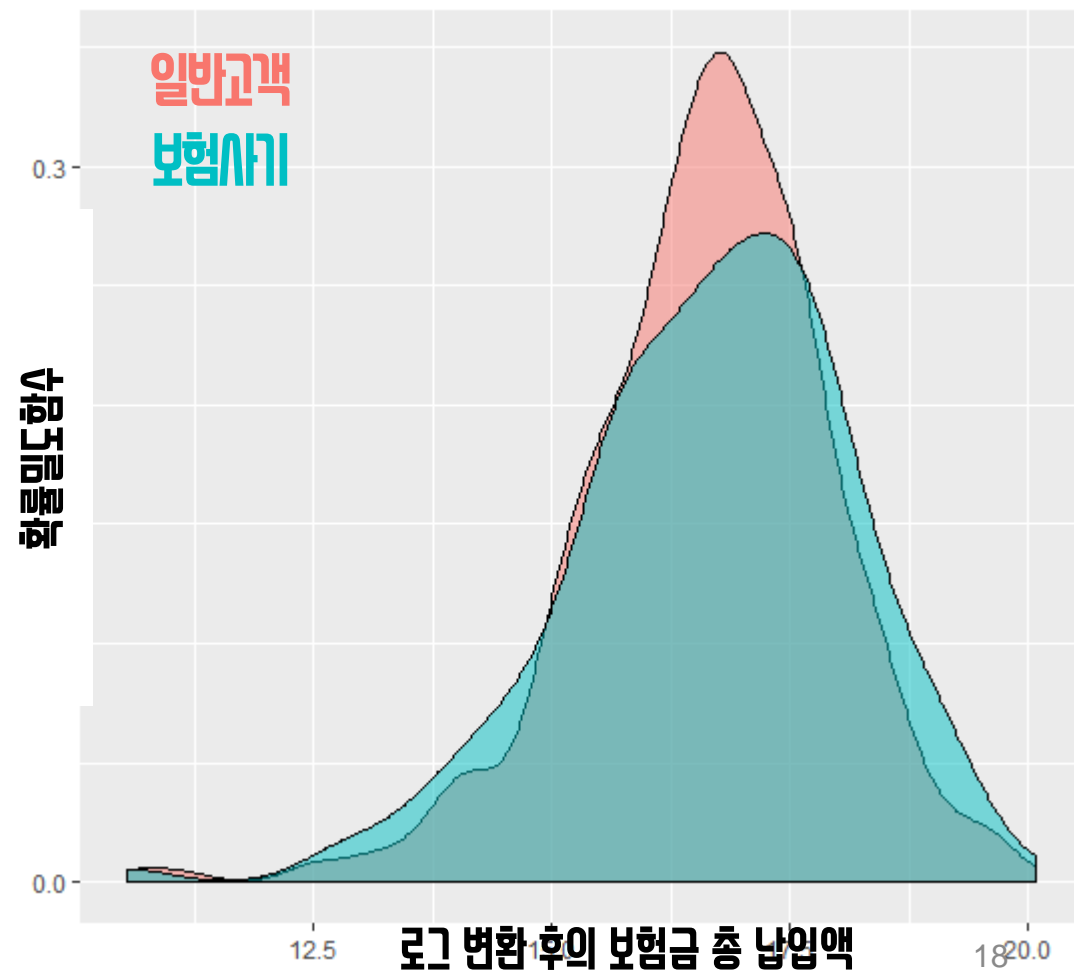
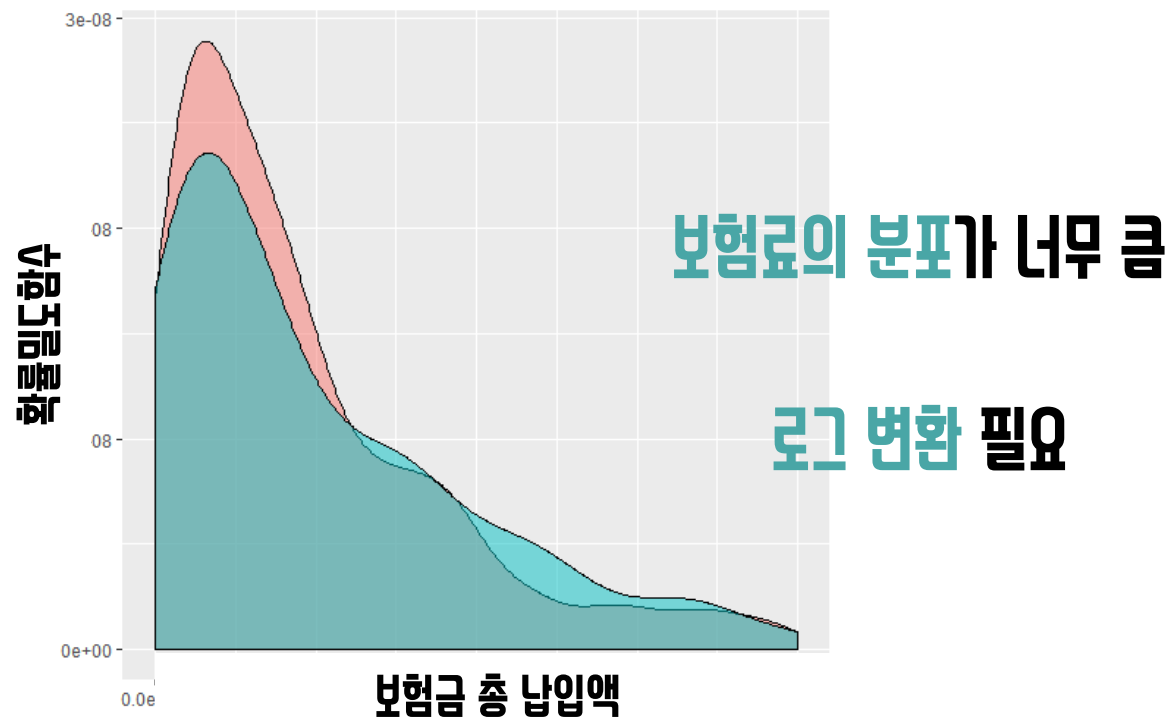
• 전직 보험 설계사

- 보험설계사 근무 경험 유무에 대한 범주형 설명 변수
- 보험설계사로 근무 했던 경험이 고객이 되어서 보험사기에 영향을 줄 수 있음
- 실제로 근무했던 사람이 보험 사기를 친 비율이 7% 높게 나왔다.



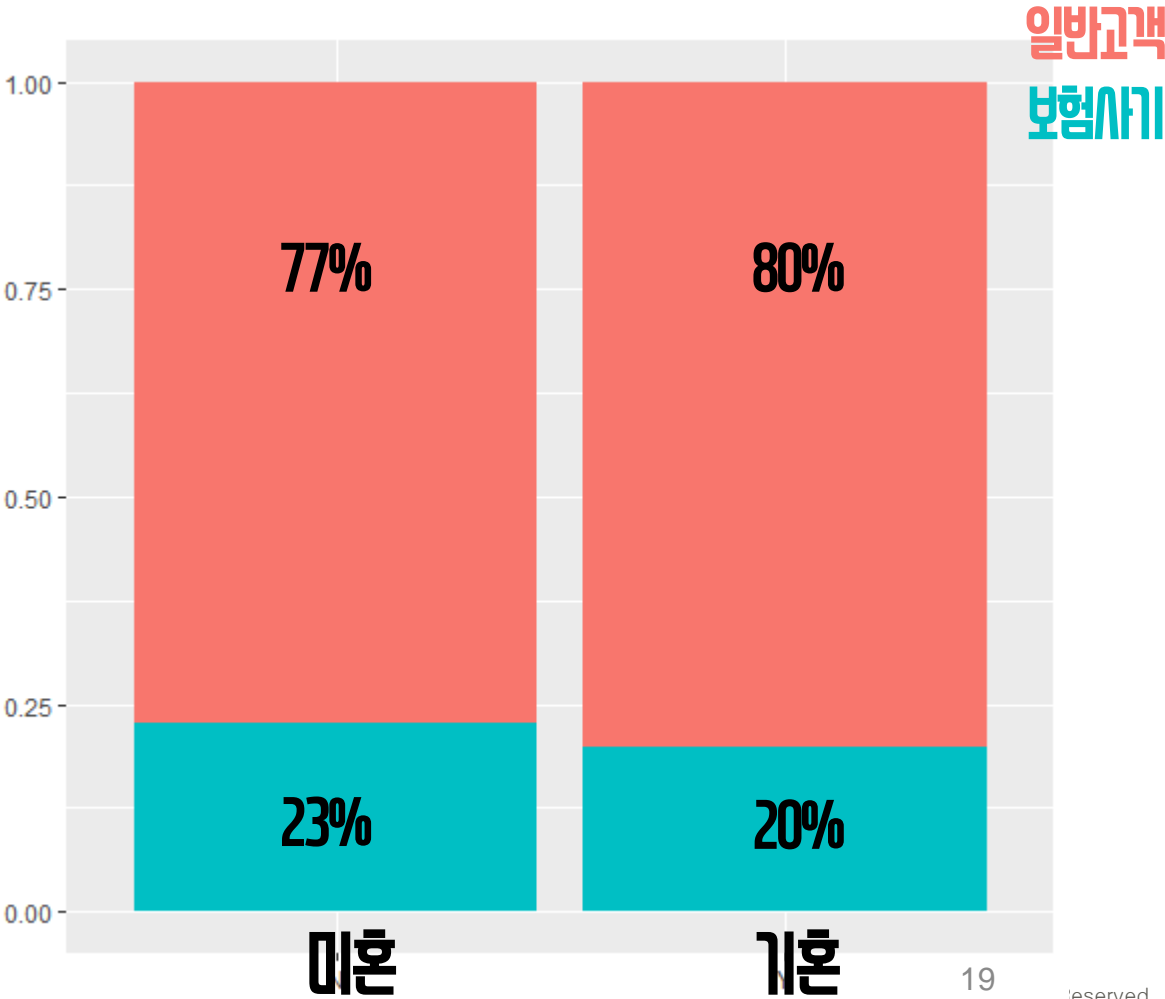
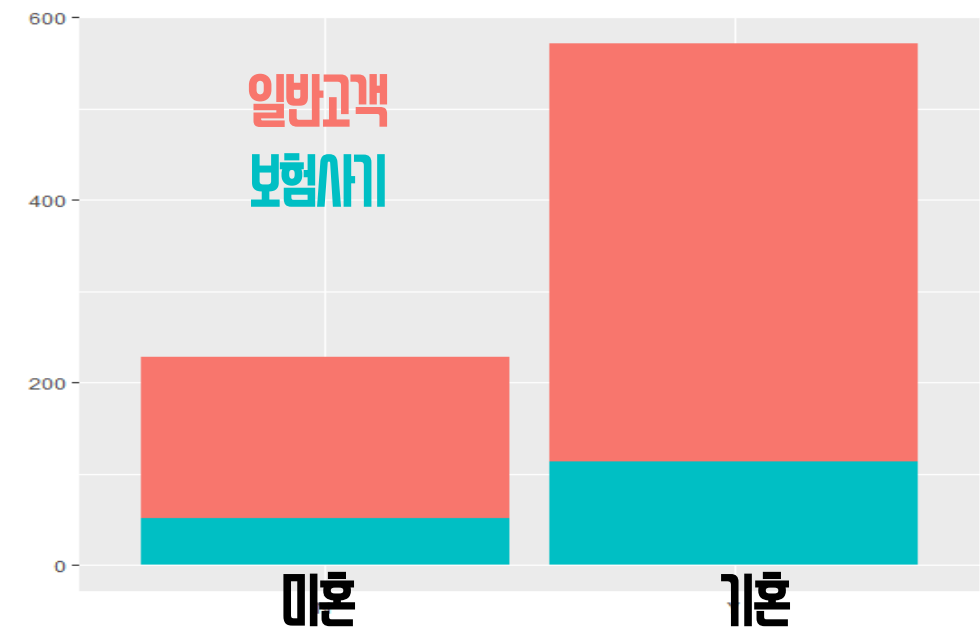
• 보험금 총 납입액

- 보험금 총 납입액이 적게는 37,400 원에서 많게는 531,643,269 원으로 다양했다.



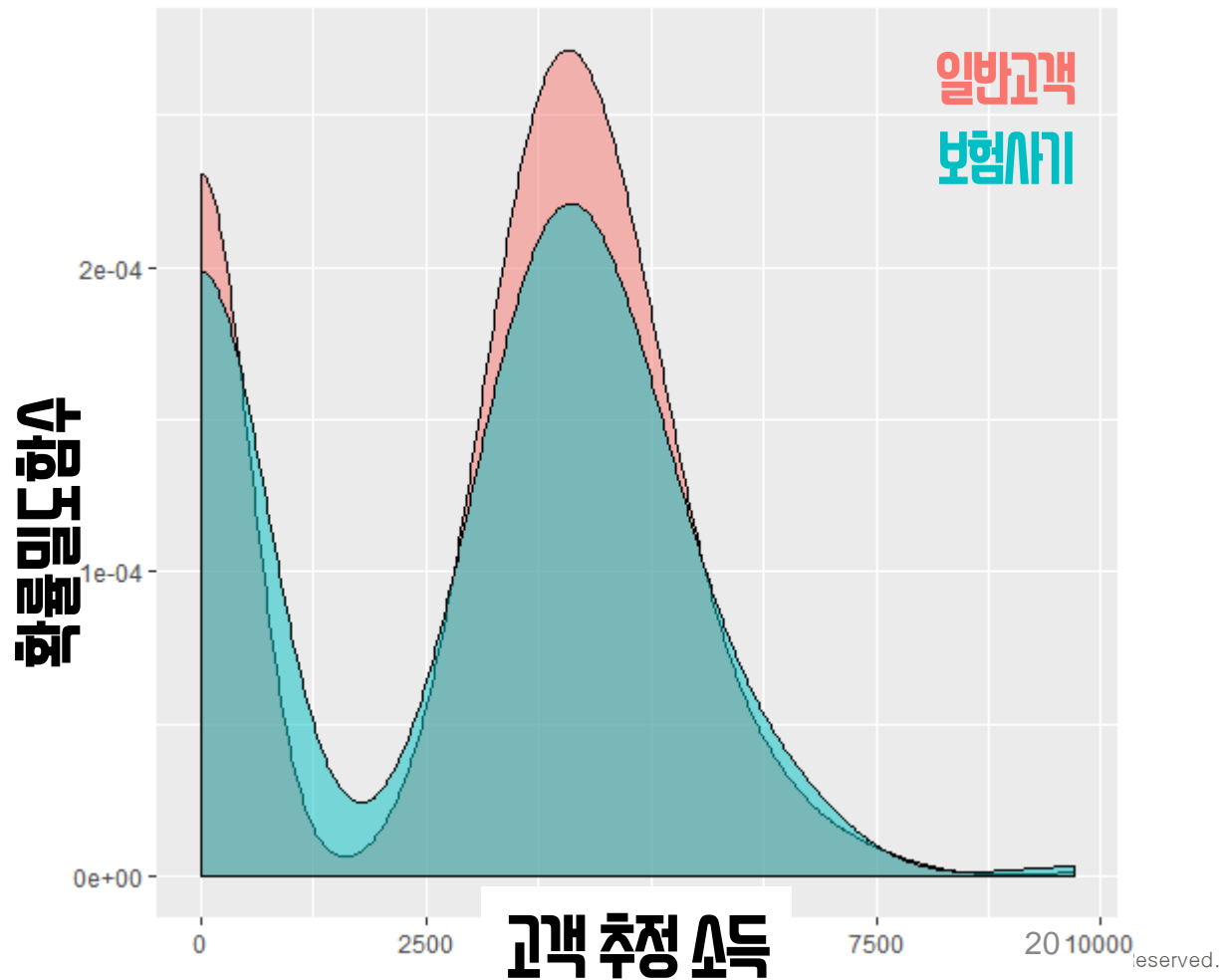
- 결혼 유무

- 결혼 유무에 대한 범주형 설명 변수
- 범주 구성 : 미혼, 기혼



• 고객 추정 소득

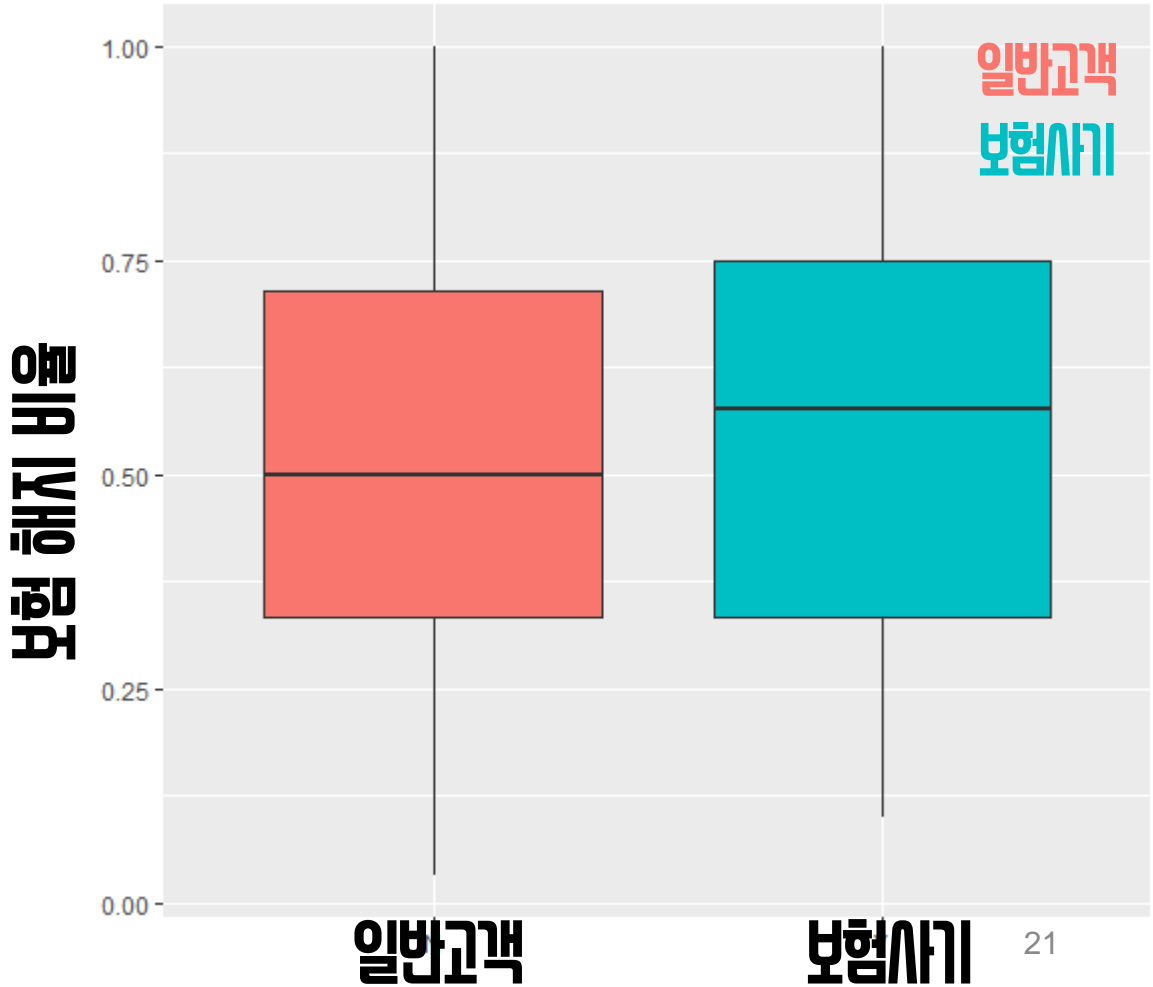
- 고객 추정 소득에 대한 연속형 설명 변수
- 소득이 없는 경우 “0” 으로 표시
- 일반고객의 경우의 “0”의 비율 : 31%
- 보험사기의 경우의 “0”의 비율 : 36%
- 보험사기여부에 상관없이 꽤 많은 수에 사람들이 소득이 없지만 보험에 가입한 것을 알 수 있었다.



- 보험 해지 비율

- $$\text{보험해지비율} = \frac{\text{해지 계약 건수}}{\text{전체 계약 건수}}$$

(0 ~ 1 사이의 범위)

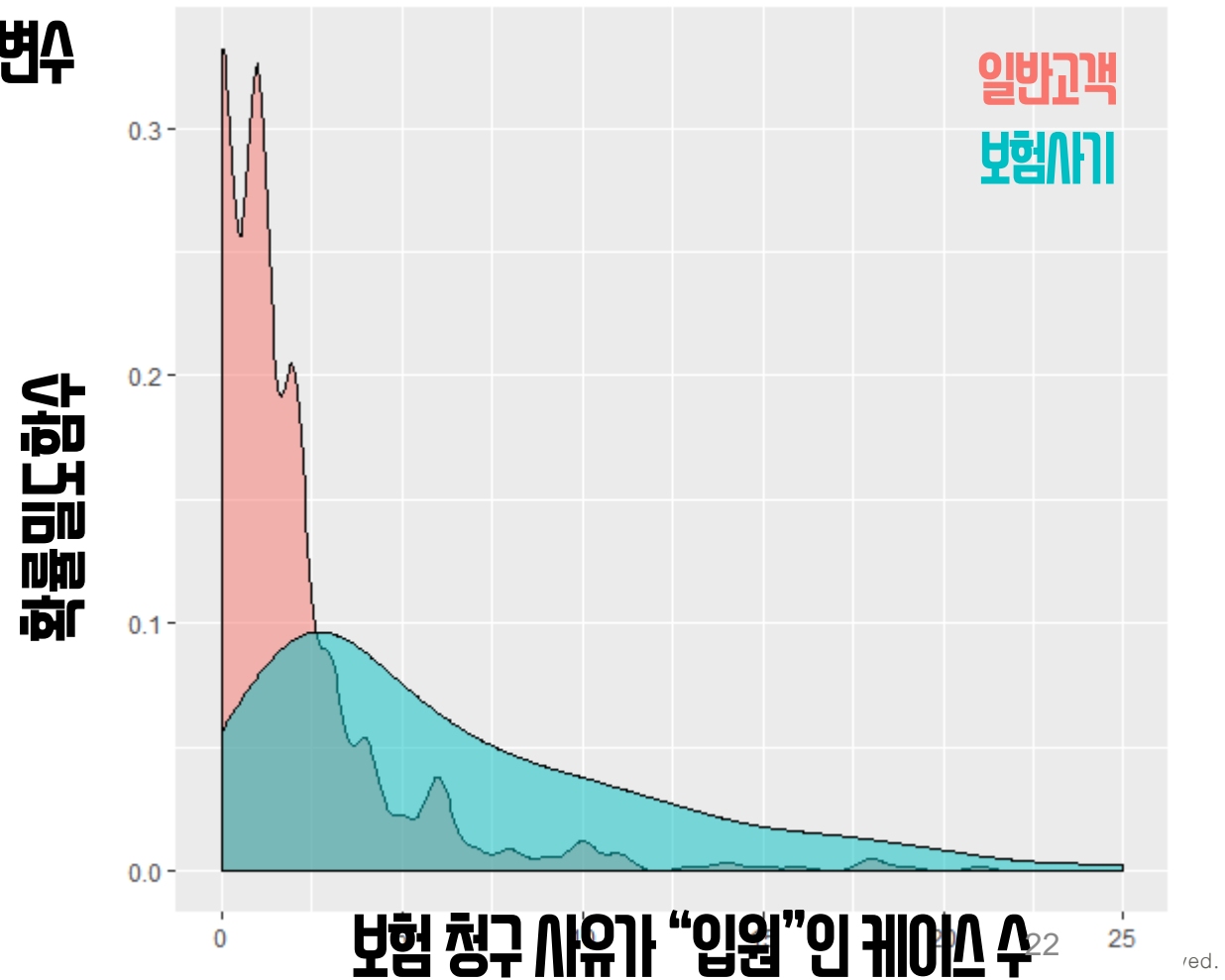


- 입원 청구 횟수

- 보험 청구 사유가 입원인 케이스의 수에 대한 연속형 설명 변수

- 일반고객과 보험사기의 분포의 차이가 뚜렷하게 나타났다.

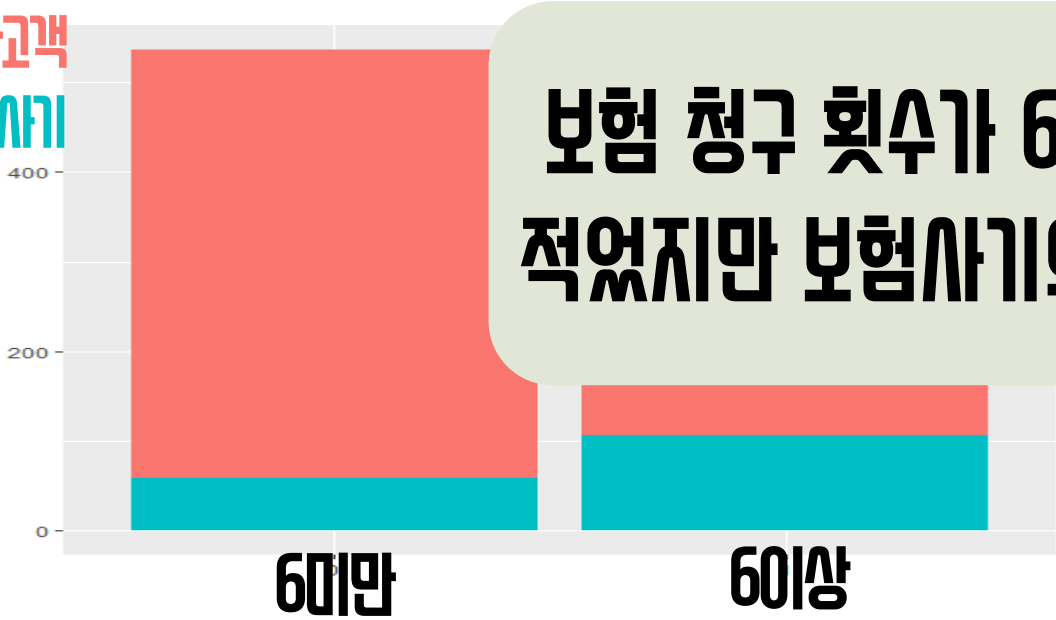
- 두 분포 모두 오른쪽으로 치우친 분포이지만, 보험사기의 경우 꼬리부분의 두께가 더 두꺼웠다.



• 보험 청구 수

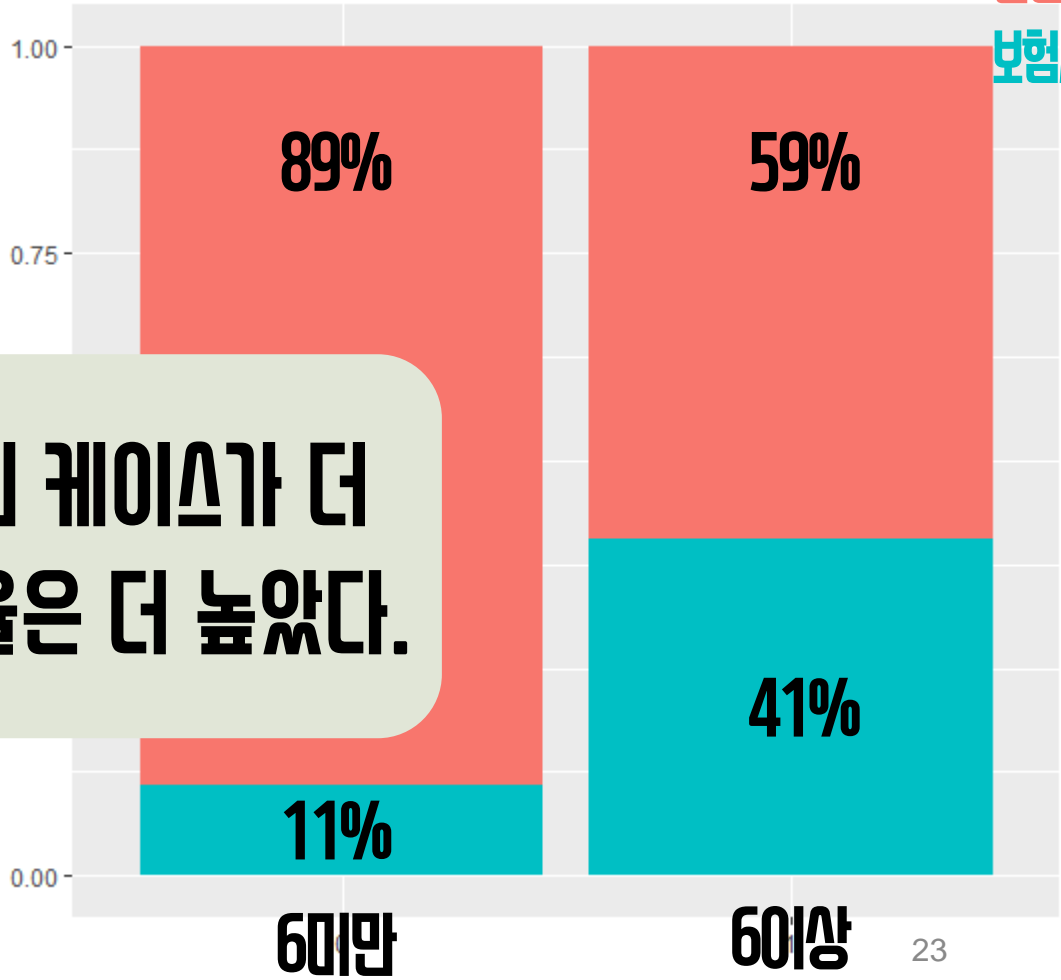
- 보험을 가입한 이후로 실제로 보험을 청구한 케이스 수
- 청구 횟수가 6이상인 경우와 6 미만인 경우로 분리

일반고객
보험사기



보험 청구 횟수가 6이상인 케이스가 더 적었지만 보험사기의 비율은 더 높았다.

일반고객
보험사기



- 계약 후 보험까지 걸리는 최소 기간

- 최초 계약일로부터 보험을 청구하였을 때까지의 걸린 최소 기간

- 최소 기간이 짧으면 짧을수록 보험사기를 의심

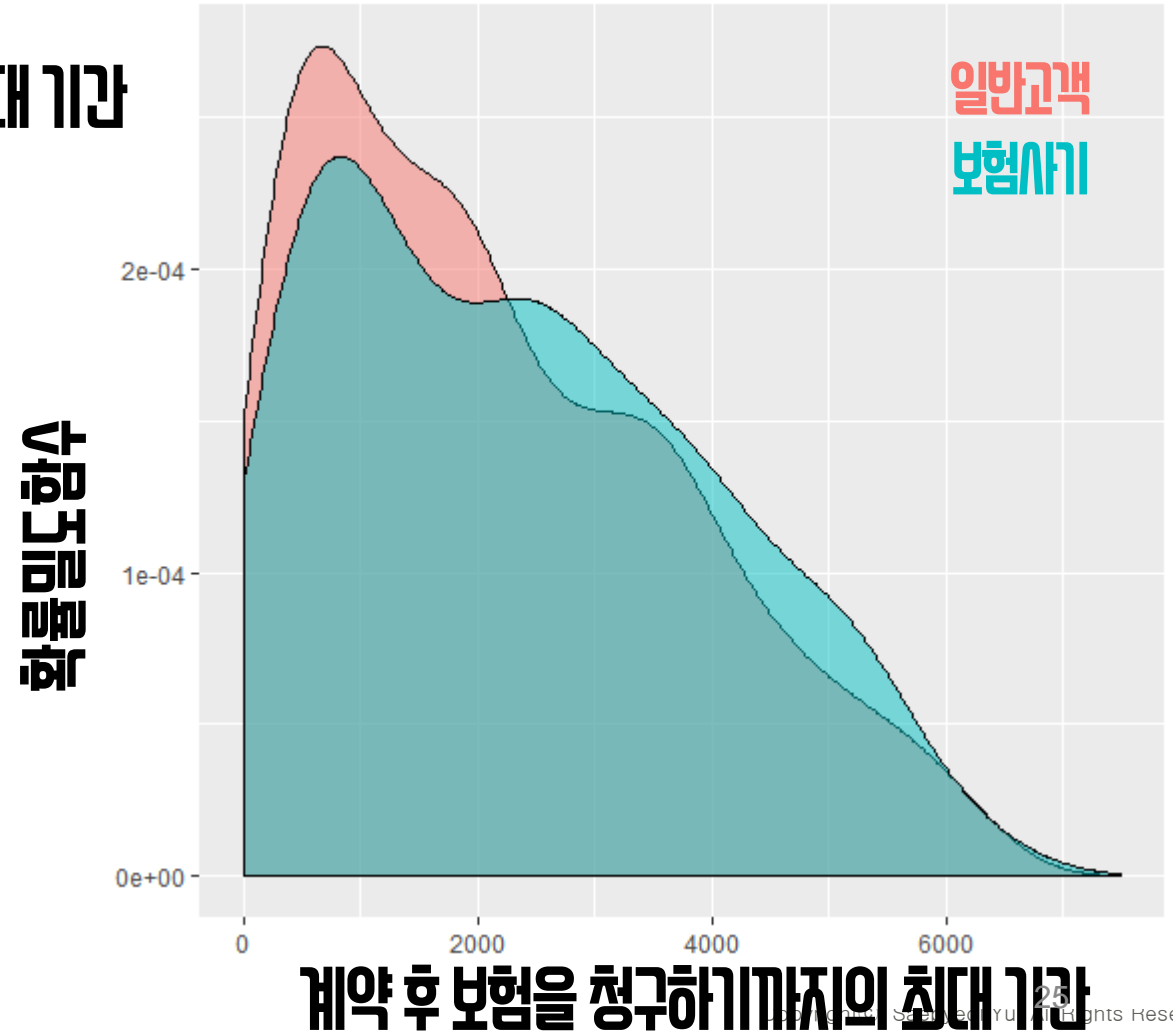
- 보험사기의 경우 기간이 짧은 쪽에 좀 더 모여있다.



계약 후 보험을 청구하기까지의 최소 기간

- 계약 후 보험까지 걸리는 최대 기간

- 최초 계약일로부터 보험을 청구하였을 때까지의 걸린 최대 기간
- 최대 기간이 길면 길수록 일반 고객일 가능성이 높음
- 일반고객의 경우 기간이 짧은 쪽에 좀 더 모여있다.



- EDA 정리

- 8개의 연속형 설명 변수 중 “보험 가입 수”, “보험금 총 납입액” 을 로그 변환
- 로지스틱 회귀 분석 시 영향력이 클 것이라 판단되는 변수 : 입원 청구 횟수, 보험 청구 횟수, 사고 원인, 병원 종류
- 로지스틱 회귀 분석 시 영향력이 작을 것이라 판단되는 변수 : 나이, 직업, 고객 추정 소득

Part 3.

로지스틱 회귀분석



- 로지스틱 회귀 모형 적합

- 15개의 설명변수를 로지스틱 회귀 모형으로 적합

설명변수	포함 여부	설명변수	포함 여부
진단서 상의 병원의 종류	0	결혼 유무	0
고객의 직업		고객 추정 소득	
사고원인	0	계약 해지 비율	
보험담당자 변경 횟수		입원 청구 수	0
보험 가입 수		보험 청구 수	
고객 나이		계약 후 최소 보험 청구 기간	
보험 설계 경험		계약 후 최대 보험 청구 기간	0
보험금 총 납입액			

- 총 5개의 변수가 유의적

- AIC = 578.1949

- BIC = 667.2025

- 변수 선택

- 검정에 의한 모형 선택 (전진선택법)

M1 : 보험사기여부 ~ 입원 청구 횟수 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 + 계약 해지 비율 + 보험담당자 변경 횟수

- 검정에 의한 모형 선택 (후진소거법)

M2 : 보험사기여부 ~ 입원 청구 횟수 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 + 결혼 유무 +
진단서 상의 병원 종류 + 보험담당자 변경 횟수

- 검정에 의한 모형 선택 (단계적선택법)

M1 : 보험사기여부 ~ 입원 청구 횟수 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 + 계약 해지 비율 + 보험담당자 변경 횟수

• 변수 선택

• AIC 기준에 의한 단계적 선택법

M3 : 보험사기여부 ~ 입원 청구 횟수 + 진단서 상의 병원 종류 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 +
결혼 유무 + 보험담당자 변경 횟수 + 계약 해지 비율 + 보험금 총 납입액

• BIC 기준에 의한 단계적 선택법

M4 : 보험사기여부 ~ 입원 청구 횟수 + 진단서 상의 병원 종류 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간

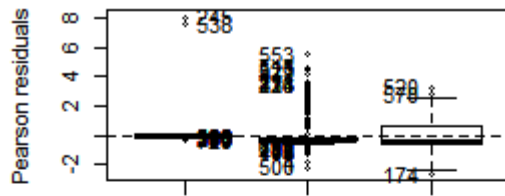
4 가지 모델 중 AIC, BIC 등을 고려하여 하나의 잠정 모형으로 선택

• **잠정 모형 선택**

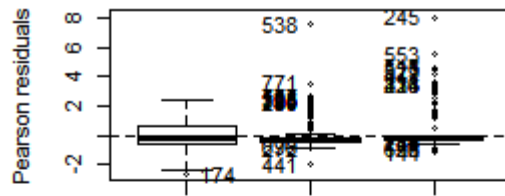
	M1	M2	M3	M4
선택 기준	검정(전진, 단계적)	검정(후진)	AIC 기준	BIC 기준
모형에 포함된 변수 개수	5개	6개	8개	4개
AIC	595.9059	568.8951	568.2238	572.1899
BIC	628.6982	611.0566	619.7545	604.9822

- M1, M2, M3, M4 모두 **AIC와 BIC가 크게 차이 나지는 않았다.**
- 따라서 “모수 절약의 원칙” 의 원칙에 따라 **M4를 잠정 모형**으로 선택한다.

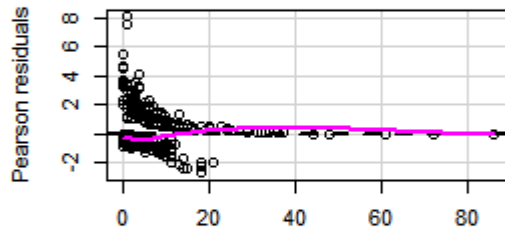
• 잠정 모형 M4의 선형 관계 확인



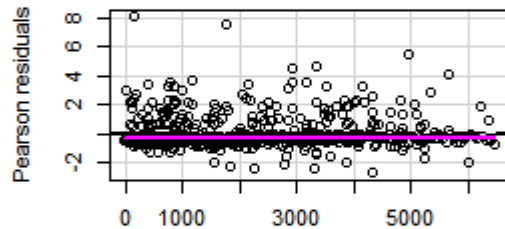
병원 종류



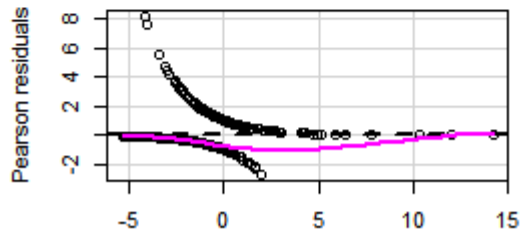
사고 원인



입원 청구 횟수



계약 후 보험 청구 최장 기간

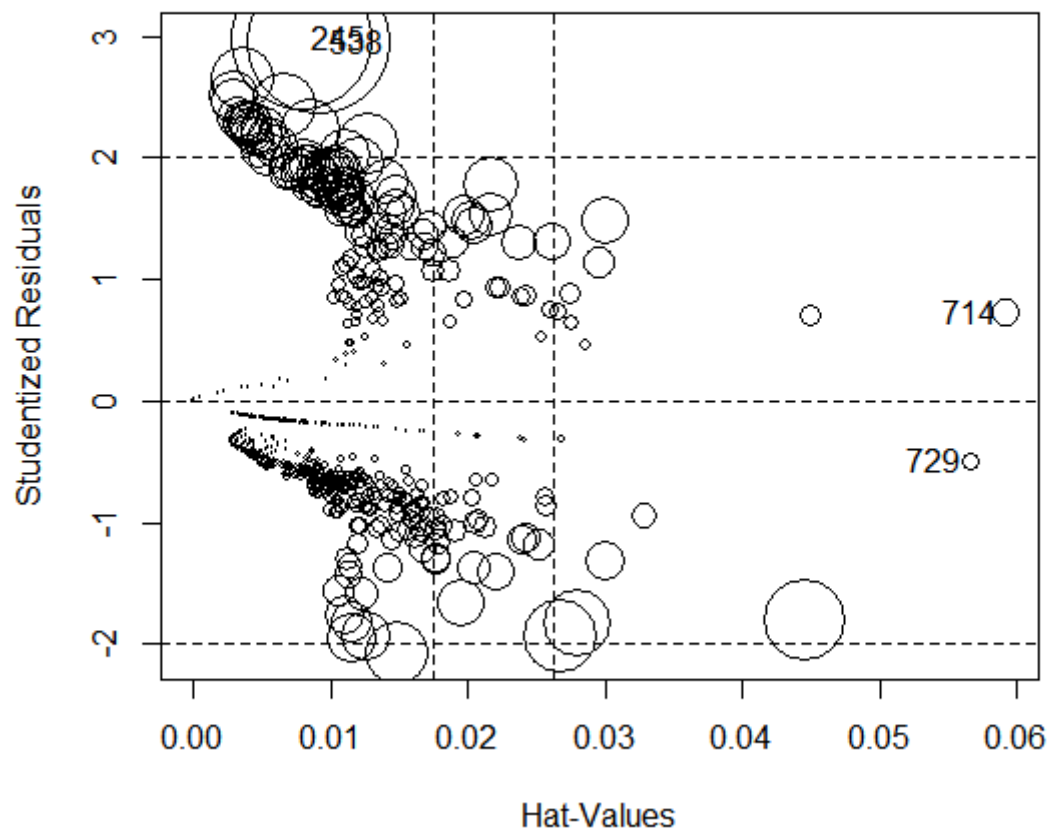


Linear Predictor

- 선형관계에서는 특이한 점을 발견하기 어려움
- 2개의 연속형 변수에 대한 2차항 추가 여부 확인

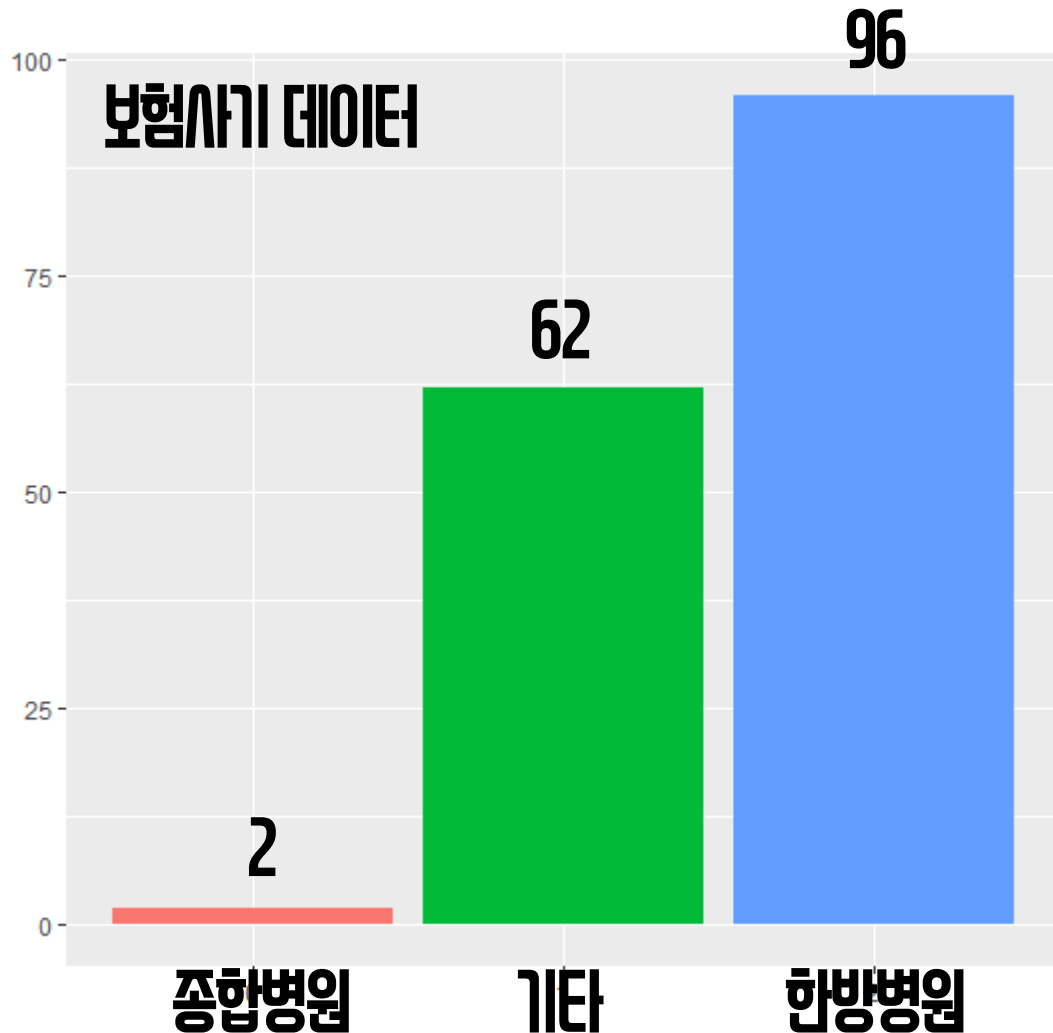
설명 변수	Test Stat	P-value
입원 청구 횟수	1.3945	0.2377
계약 후 보험 청구 최장 기간	0.2238	0.6361

- 잠정 모형 M4의 이상값 탐지



- 스튜던트화 잔차에서 +3과 -3을 벗어나는 관찰값
→ 245번, 538번
- Cook' Distance에서 상대적으로 큰 원을 가진 관찰값
→ 245번, 538번
- Hat-Values 기준으로 벗어나는 관찰값
→ 없음

- 잠정 모형 M4의 이상값 확인



- 이상값 후보인 245번, 538번이 종합병원에 해당
- 보험사기의 경우 진단서를 발급받은 대부분의 병원이 기타 혹은 한방병원이다
- 종합병원의 경우 단 2개의 케이스만 존재
- 모형에 미치는 영향력이 크다고 생각하여 2개 이상값 제거

- 변수 선택

- 검정에 의한 모형 선택 (전진선택법)

M5 : 보험사기여부 ~ 입원 청구 횟수 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 + 계약 해지 비율 + 보험담당자 변경 횟수

- 검정에 의한 모형 선택 (후진소개법)

M6 : 보험사기여부 ~ 입원 청구 횟수 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 + 결혼 유무 +
진단서 상의 병원 종류 + 보험담당자 변경 횟수

- 검정에 의한 모형 선택 (단계적선택법)

M5 : 보험사기여부 ~ 입원 청구 횟수 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 + 계약 해지 비율 + 보험담당자 변경 횟수

• 변수 선택

• AIC 기준에 의한 단계적 선택법

M7 : 보험사기여부 ~ 입원 청구 횟수 + 진단서 상의 병원 종류 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 +
결혼 유무 + 보험담당자 변경 횟수 + 계약 해지 비율 + **고객 추정 소득**

• AIC 기준에 의한 단계적 선택법 (이상값 제거 전)

M3 : 보험사기여부 ~ 입원 청구 횟수 + 진단서 상의 병원 종류 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간 +
결혼 유무 + 보험 담당자 변경 횟수 + 계약 해지 비율 + **보험금 총 납입액**

• BIC 기준에 의한 단계적 선택법

M8 : 보험사기여부 ~ 입원 청구 횟수 + 진단서 상의 병원 종류 + 사고 원인 + 계약 후 보험 청구까지의 최대 기간

4 가지 모델 중 AIC, BIC 등을 고려하여 하나의 잠정 모형으로 선택

• 잠정 모형 선택

	M5	M6	M7	M8
선택 기준	검정(전진, 단계적)	검정(후진)	AIC 기준	BIC 기준
모형에 포함된 변수 개수	5개	6개	8개	4개
AIC	586.1883	548.9123	548.4762	551.6391
BIC	618.963	591.0513	599.9794	584.4139

- M6, M7, M8의 경우 AIC와 BIC가 크게 차이 나지는 않았다.
- 따라서 “모수 절약의 원칙” 의 원칙에 따라 M8를 잠정 모형으로 선택한다.

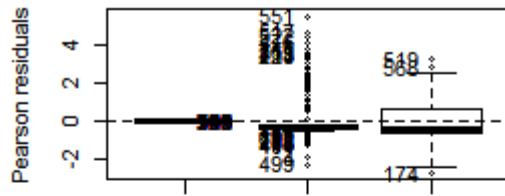
• 이상값 제거 후 AIC와 BIC의 변화

	M1	M2	M3	M4
선택기준	검정(전진, 단계적)	검정(후진)	StepAIC	StepBIC
모형에 포함된 변수 개수	5개	6개	8개	4개
AIC	595.9059	568.8951	568.2238	572.1899
BIC	628.6982	611.0566	619.7545	604.9822

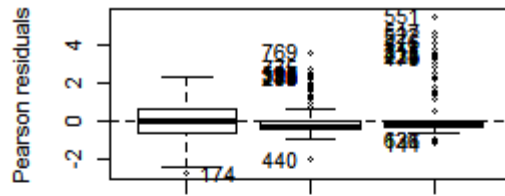
	M5	M6	M7	M8
선택기준	검정(전진, 단계적)	검정(후진)	StepAIC	StepBIC
모형에 포함된 변수 개수	5개	6개	8개	4개
AIC	586.1883	548.9123	548.4762	551.6391
BIC	618.963	591.0513	599.9794	584.4139

- 포함된 변수가 달라진 M3 와 M7을 제외하고 모두 AIC 와 BIC 값이 하락하였다.

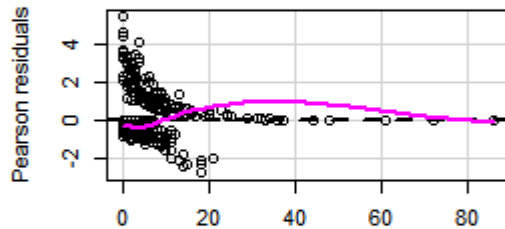
• 잠정 모형 M8의 선형 관계 확인



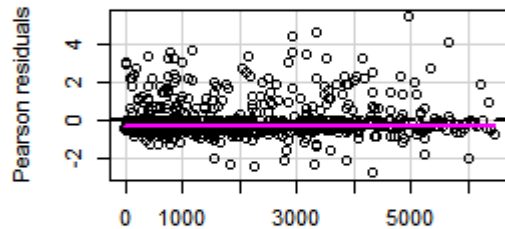
병원 종류



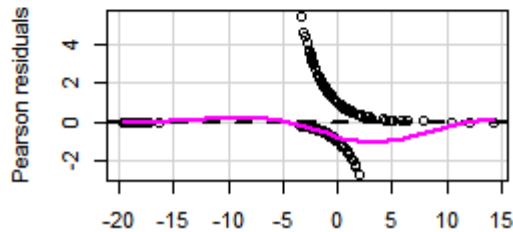
사고 원인



입원 청구 횟수



계약 후 보험 청구 최장 기간



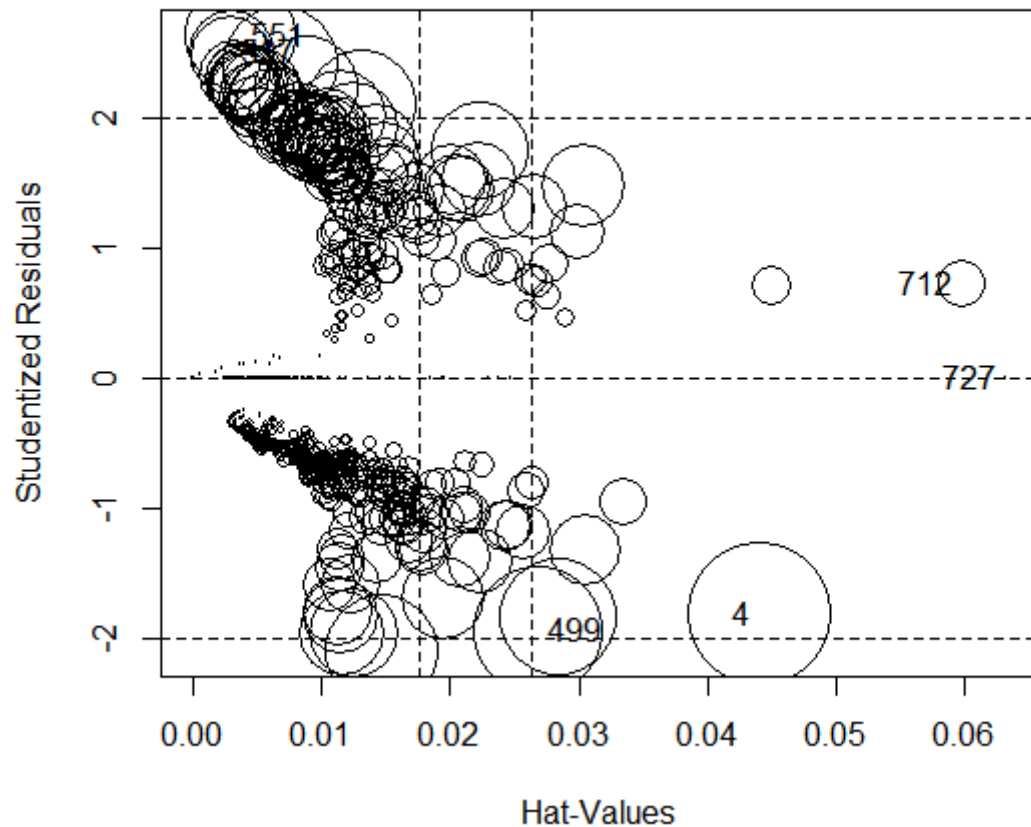
Linear Predictor

• 선형관계에서는 특이한 점을 발견하기 어려움

• 2개의 연속형 변수에 대한 2차항 추가 여부 확인

설명 변수	Test Stat	P-value
입원 청구 횟수	1.4427	0.2297
계약 후 보험 청구 최장 기간	0.2227	0.6370

- 잠정 모형 M8의 이상값 탐지



- 스튜던트화 잔차에서 +3과 -3을 벗어나는 관찰값
→ 없음
- Hat-Values 기준으로 벗어나는 관찰값
→ 없음
- Cook' Distance에서 상대적으로 큰 원을 가진 관찰값
→ 499번, 4번
- 심각한 이상값은 없음.

• 다중공선성의 문제

- 설명 변수들 사이에 강한 선형 관계가 존재하는 것이 다중공선성의 문제
- 다중공선성은 회귀 모형의 가정과 직접적인 연관이 있는 것은 아니지만, 회귀 모형의 추정 결과를 해석하는 과정에 큰 영향을 미칠 수 있는 문제가 된다.
- 모형에 포함된 변수 4개 모두 VIF 값이 약 1이므로 4개의 설명변수 사이에 다중공선성의 문제는 없는 것으로 보임.

설명 변수	VIF
진단서 상의 병원 종류	1.019776
사고 원인	1.034339
입원 청구 횟수	1.144405
계약 후 보험 청구 최대 기간	1.068714

- 최종 모형

$$\log \frac{\text{보험사기일 확률}}{1 - \text{보험사기일 확률}} = -17.55 + 16.31 \text{ 병원 종류_기타} + 17.2 \text{ 병원 종류_한방병원} - 0.71 \text{ 사고원인_비스} \\ - 1.14 \text{ 사고원인_낮음} + 0.18 \text{ 입원 청구 횟수} - 0.0002 \text{ 계약 후 보험 청구 최대 기간}$$

$$\frac{\text{보험사기일 확률}}{1 - \text{보험사기일 확률}} = e^{-17.55 + 16.31 \text{ 병원 종류_기타} + 17.2 \text{ 병원 종류_한방병원} - 0.71 \text{ 사고원인_비스} \\ - 1.14 \text{ 사고원인_낮음} + 0.18 \text{ 입원 청구 횟수} - 0.0002 \text{ 계약 후 보험 청구 최대 기간}}$$

• Odds Ratio에 의한 효과 분석

• 로지스틱 회귀 모형 및 회귀 계수

절편	병원 종류_기타	병원 종류_한방병원	사고원인_비상	사고원인_낮음	입원 청구 횟수	계약 후 최대 기간
-17.55	16.31	17.2	-0.71	-1.14	0.18	-0.0002

• 지수변환으로 Odds에 대한 모형으로 변환

절편	병원 종류_기타	병원 종류_한방병원	사고원인_비상	사고원인_낮음	입원 청구 횟수	계약 후 최대 기간
0.000000002	12,169,940	29,596,120	0.4922364	0.3186669	1.196933	0.9997941

• 대략적인 해석

→ 다른 설명변수의 수준은 고정

→ 1보다 크면 보험 사기일 Odds Ratio 증가, 1보다 작으면 보험 사기일 Odds Ratio 감소

• Odds Ratio에 의한 효과 분석

- 병원 종류가 “기타” 인 경우에는 병원 종류가 “종합병원”일 경우에 비해 보험사기의 Odds ratio는..?

→ $\exp(\hat{\beta}_1) = \exp(16.31) = 12,169,940$ 배 증가

→ $100 * (12,169,940 - 1) = 1,216,993,900$ % 증가

- 병원 종류가 “한방병원” 인 경우에는 병원 종류가 “종합병원”일 경우에 비해 보험사기의 Odds ratio는..?

→ $\exp(\hat{\beta}_2) = \exp(17.2) = 29,596,120$ 배 증가

→ $100 * (29,596,120 - 1) = 2,959,611,900$ % 증가

• Odds Ratio에 의한 효과 분석

- 사고 원인이 “보통” 인 경우에는 사고 원인이 “높음”일 경우에 비해 보험사기의 Odds ratio는..?

→ $\exp(\hat{\beta}_3) = \exp(-0.71) = 0.49$ 배 감소

→ $100 * (0.49 - 1) = 51\%$ 감소

- 사고 원인이 “낮음” 인 경우에는 사고 원인이 “높음”일 경우에 비해 보험사기의 Odds ratio는..?

→ $\exp(\hat{\beta}_4) = \exp(-1.14) = 0.32$ 배 감소

→ $100 * (0.32 - 1) = 68\%$ 감소

보험 사기인 경우 사고 원인이
어느 정도 정해져 있다.

• Odds Ratio에 의한 효과 분석

- 입원 청구 횟수가 1건 늘어난다면 보험사기의 Odds ratio는..?

$$\rightarrow \exp(\hat{\beta}_5) = \exp(0.17) = 1.19 \text{ 배 증가}$$

$$\rightarrow 100 * (1.19 - 1) = 19 \% \text{ 증가}$$

- 입원 청구 횟수가 10건 늘어난다면 보험사기의 Odds ratio는..?

$$\rightarrow \exp(10 * \hat{\beta}_5) = \exp(1.7) = 5.47 \text{ 배 증가}$$

$$\rightarrow 100 * (5.47 - 1) = 447 \% \text{ 증가}$$

**입원 청구 횟수가 늘어날수록
보험 사기일 확률도 높아진다.**

- Odds Ratio에 의한 효과 분석

- 계약 후 보험 청구하기까지의 최대 기간이 1일 늘어난다면 보험사기의 Odds ratio는..?

→ $\exp(\hat{\beta}_6) = \exp(-0.0002) = 0.99$ 배 감소

→ $100 * (0.99 - 1) = 1\%$ 감소

- 계약 후 보험 청구하기까지의 최대 기간이 365일 늘어난다면 보험사기의 Odds ratio는..?

→ $\exp(365 * \hat{\beta}_6) = \exp(-0.0002) = 0.92$ 배 감소

→ $100 * (0.92 - 1) = 8\%$ 감소

계약 후 보험 청구를 늦게 할수록
보험 사기일 확률이 감소한다

Part 4.

예측 및 결론



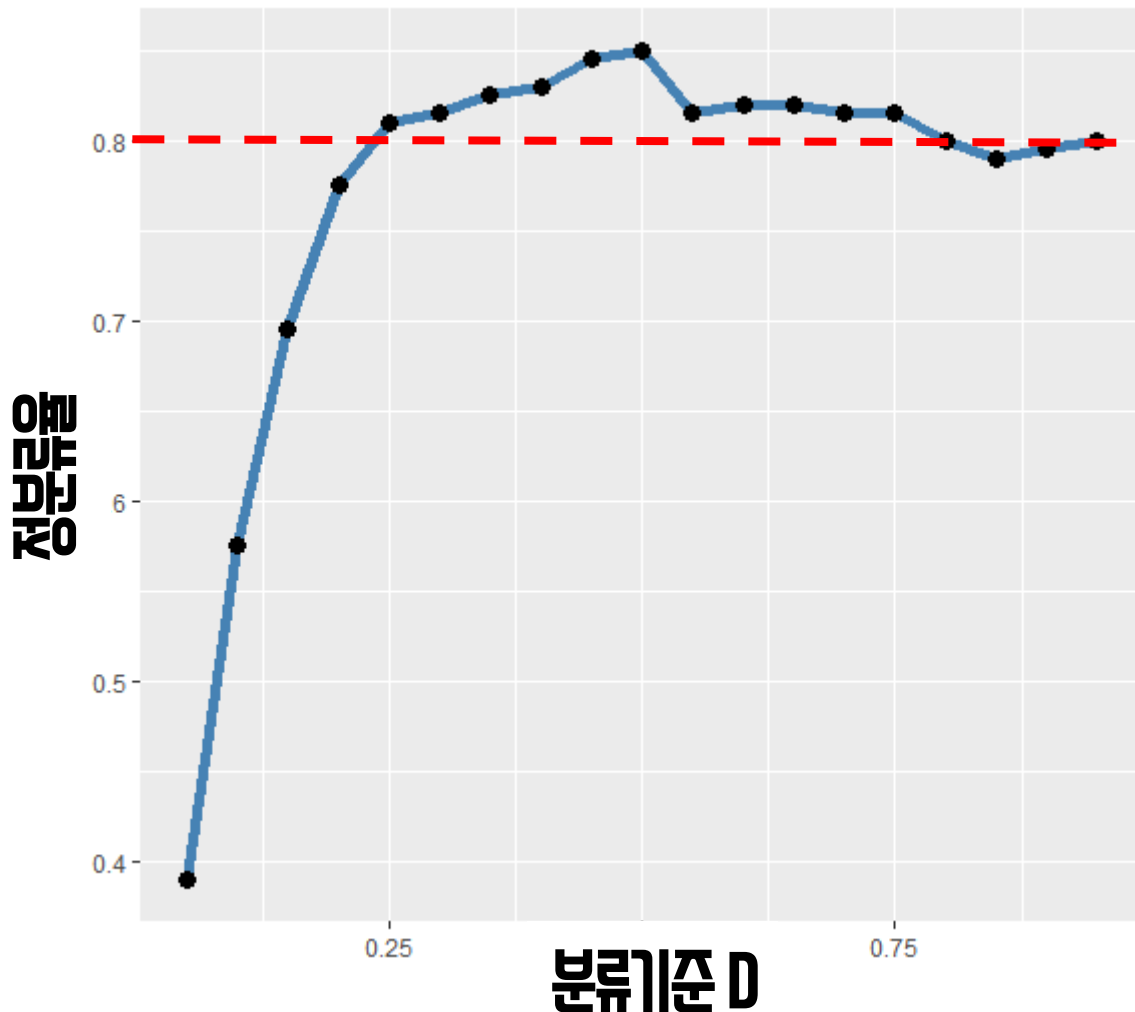
• 예측 결과 확인

Test Data (기준 0.5)		예측	
		일반고객	보험사기
실제	일반고객	155	5
	보험사기	25	15

- 정분류율 : 85%
- 수정된 정분류율 : 5%

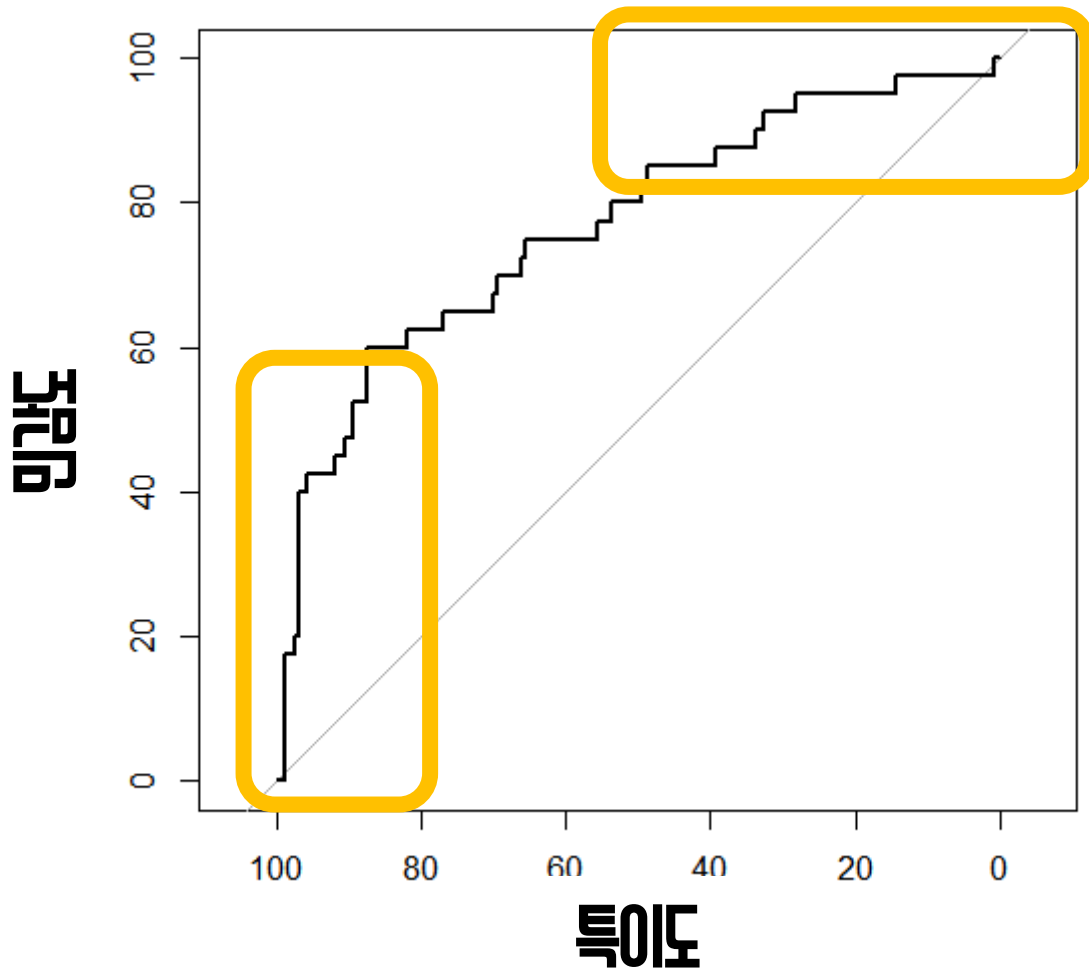
- Test 데이터에 포함된 200개의 데이터 중 올바르게 분류한 케이스는 170개이다.
- 40건의 보험사기 중 15건을 실제 보험사기라고 예측하였다.

• 정분류율의 변화



- 분류 기준인 D 의 값이 0.25 ~ 0.75 사이인 경우 정분류율이 80%를 넘었으며, 수정된 정분류율 역시 이보다 크다.
- 정분류율이 가장 높은 경우는 $D=0.5$ 일 때 85%
- 정분류율이 가장 낮은 경우는 $D=0.05$ 일 때 39%
- 정분류율이 가장 높은 $D=0.5$ 를 선택

• ROC Curve



- Area Under curve : 76.76%
- 특이도가 100%에서 80%로 떨어지는 동안 민감도는 0%에서 40%로 급상승
- 특이도가 40%에서 0%로 떨어지는 동안 민감도는 80%에서 100%로 천천히 상승

• 결론

- 보험사기인 경우 진단서를 “종합병원”이 아닌 “한방병원”이나 “기타”에서 발급받아 제출하는 경향이 있음.
- 보험사기인 경우 주로 발생하는 사고 원인(질병 코드)가 어느정도 존재.
- 입원할 필요가 없지만 고의로 입원하거나, 입원을 하지 않은 경우에도 입원으로 보험을 청구한 경우가 많음.
- 고의적으로 보험 사기를 계획하고 보험을 가입하는 경우도 많음.
- 따라서 최초 계약일로 부터 보험을 청구한 기간이 길면 길수록 일반고객일 확률이 높음.

• 결론

- 보험사기를 예측하는 로지스틱 회귀 모형에 “진단서 발급 병원 종류”, “사고 원인”, “입원 청구 횟수”, “계약 후 보험 청구 최대 기간”의 4개의 변수가 포함.
- 로지스틱 모형이 없었다면 전체 데이터 중 20%에 해당하는 보험사기를 예측할 수 없었음.
- 전체 보험 사기(40 obs) 중 37.5%(15 obs)를 예측.
- 보험사기를 20% 에서 12.5%로 줄일 수 있음.

- 한계점

- 보험사기의 비율이 20%로 많은 편은 아니었기 때문에 좀 더 확실한 모델을 만들지 못한 아쉬움이 있다.
- 종합병원의 진단서를 가지고 사기를 친 케이스가 Train data 안에는 단 2건에 불과했다.
 - 따라서 모델을 세울 때 이상값으로 제거하게 되었다.
 - 그러나 종합병원이라고 아예 보험사기가 없는 것은 아닐 것
 - 교차검증(Cross-Validation) 을 통해 전체 데이터 셋에 대한 모델을 세운다면 해결 가능

- 한계점

- 사고 원인에 따라 보험 사기 비율이 뚜렷하게 나타났기 때문에
높은 사고 원인 과 비슷한 사고 원인, 낮은 사고 원인 별로 모델을 각각 세운다면 좀 더 높은 예측율을 보일 것이다.
- 로지스틱 모형 외에도, SVM, XGBOOST, RANDOM FOREST, ANN, CNN 등
좀 더 다양한 알고리즘을 적용하여 비교함으로써, 최적의 알고리즘을 찾는 시도가 필요해 보인다.

감사합니다